Intensive Care Med (2026) 52:11–21
https://doi.org/10.1007/s00134-025-08217-0

ORIGINAL

Individualised treatment effects of enhanced early mobilisation in mechanically ventilated patients: a secondary analysis of the TEAM trial

Carol L. Hodgson[1,2,3,4*], Alexandra B. Spicer[5], Tessa Broadley[1], Michael Bailey[1], Rinaldo Bellomo[1,2,6,7], Kathy Brickell[8], Heidi Buhr[9], Belinda J. Gabbe[10], Doug W. Gould[11], Meg Harrold[12,13], Alisa M. Higgins[1,4], Sally Hurford[14], Theodore J. Iwashyna[15], Ary Serpa Neto[1,6], Alistair D. Nichol[1,8,16], Jeffrey J. Presneill[7,17], Stefan J. Schaller[18,19], Janani Sivasuthan[1], Claire J. Tipping[3], Steven Webb[1,20], Emma J. Graham Linck[5], Pratik Sinha[21,22], Matthew W. Semler[23,24], Paul J. Young[1,2,14,25] and Matthew M. Churpek[5]

Abstract

Purpose: Benefit or harm from early mobilisation (EM) in mechanically ventilated patients may vary by individual patient characteristics. We used machine learning to predict individualised treatment effects (ITEs) in the "Early Active Mobilization during Mechanical Ventilation in the ICU" (TEAM) trial.

Methods: This was a secondary analysis of the TEAM trial using a causal inference approach to estimate ITEs, which compared enhanced EM to usual care EM. Baseline variables in the original publication were used as predictor variables. The primary outcome was death by day 180. The dataset was randomly split into two halves (train and test) by site. In the training data, fivefold cross-validation was used to compare six candidate machine learning algorithms. The best-performing model was evaluated in the test dataset. Patients were stratified into tertiles based on predicted ITEs, reflecting estimated benefit, no effect or harm.

Results: We included 687 patients from 40 sites, and 141 (20.5%) patients died by day 180. Predicted ITEs in the test cohort ranged from an absolute 34.0% reduction to a 39.3% increase in mortality with enhanced EM. The interaction term between the model predictions and treatment assignment demonstrated significant heterogeneity of treatment effect ($p = 0.006$). Patients predicted to respond poorly to enhanced EM therapy were more likely to receive vasopressors, have diabetes and have lower RASS scores at baseline, compared to patients predicted to have benefit.

Conclusion: Using baseline characteristics, a machine learning model identified patients with estimated benefit or harm with enhanced EM. Future testing of a personalised approach to mobilisation in the ICU is warranted.

Keywords: Early mobilisation, Rehabilitation, Individualised treatment effect, Clinical trial, Intensive care

Introduction

Critical illness and prolonged intensive care unit (ICU) stays are associated with complications, including muscle weakness, functional impairments and prolonged recovery [1, 2]. One of the most debilitating outcomes of prolonged ICU admission is ICU-acquired weakness, which can severely limit recovery, quality of life and survival [3–5]. Early rehabilitation and mobilisation (EM)

*Correspondence: carol.hodgson@monash.edu
[1] Australian and New Zealand Intensive Care Research Centre, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia
Full author information is available at the end of the article

Springer

have been proposed to mitigate these negative effects, with the goal of promoting functional recovery and reducing long-term disability [6, 7]. However, despite the widespread use of EM protocols in ICU settings, the optimal timing, intensity and patient characteristics that predict benefit from EM remain unclear [8, 9].

Randomised clinical trials (RCTs) have demonstrated that EM can increase muscle strength and functional recovery, [7, 10, 11] but these studies typically report average treatment effects (ATE), which may obscure variations in individual responses to the intervention. Subgroup analyses of these trials often fail to account for the complex interplay of multiple patient characteristics that may influence the effectiveness of EM [12]. For example, factors such as age, comorbidities, sedation levels and the severity of illness may all modulate how a patient responds to EM [13, 14]. These conventional subgroup approaches are limited by their reliance on predefined categories, low statistical power and inability to model continuous or interacting variables, making them insufficient for capturing the full spectrum of heterogeneity of treatment effect (HTE). HTE refers to the variation in treatment response across individuals, and understanding it is essential for tailoring interventions to those most likely to benefit. To better understand these nuances, machine learning techniques offer a promising solution by accounting for multiple factors simultaneously [15]. In RCTs, these machine learning models have identified patients predicted to benefit from and be harmed by different interventions, such as oxygen targets [16] and intubation using a bougie [17]. Understanding the HTE may allow for prospective testing of more personalised, evidence-based approaches to ICU mobilisation, ultimately improving patient outcomes and reducing unnecessary risks [18].

The aim of this study was to use machine learning methods to examine how individual patient characteristics influence the effect of enhanced versus usual care EM on outcomes in the "Early Active Mobilisation during Mechanical Ventilation in the ICU" (TEAM) RCT [19].

Methods

Study population

Data for this secondary analysis compared enhanced EM ($N=346$) to usual care EM ($N=341$) for adult patients undergoing invasive mechanical ventilation in the ICU in the TEAM RCT. This multicentre trial enrolled patients from 49 different hospitals across 6 countries between February 2018 and November 2021. Patients in the enhanced EM group received daily physiotherapy sessions aimed at attaining the highest level of mobilisation for a longer duration of time (that was safe and tolerable) for the patient for an average of 21 min per day, up to 7 days a week. In contrast, those in the usual care EM group were mobilised according to the standard protocol followed at each site which was an average of 9 min per day for 5 days a week.

The TEAM trial was approved by the institutional review boards of each participating country, and a waiver of informed consent was approved for the data to be transferred to the University of Wisconsin-Madison for analysis. The United Kingdom's consent form did not permit data transfer to a third country, resulting in its omission from this analysis.

Statistical analysis

Study design

This study investigated the heterogeneity of treatment effect using an effect-based analysis (eFigure 1) and followed guidelines from the Predictive Approaches to Treatment Effect Heterogeneity (PATH) statement [20]. The TEAM trial dataset was randomly split into two halves (train and test) by site to ensure rigorous external validation in sites not included in the training data.

Predictor variables

All the baseline variables listed in Table 1 and Tables S1–S5 of the original publication were candidates for inclusion. Variables with high missingness rates (> 35%) or high correlation (Pearson correlation coefficient > |0.7|) were removed. Small categories (< 10%) were combined or dropped if their combination with another variable was not clinically meaningful. The final variable list included demographics, frailty and function scores, conditions listed in the Functional Comorbidity Index, characteristics of the ICU admission, pre-randomisation treatment and selected vital signs and laboratory results (eTable 1). These characteristics were compared across the training and testing datasets using descriptive statistics (eTable 2). Bagged trees were built in the training data using the caret package's 'preProcess' function in R to impute missing values in both the training and testing data [21].

Primary outcome

The primary outcome for this analysis was 180-day mortality, a pre-specified secondary outcome of the TEAM trial. This approach maintains comparability with the primary trial and aligns with its predefined statistical analysis plan.

Model derivation

In the training data, fivefold cross-validation was used to compare six candidate machine learning algorithms that were chosen based on our prior work [16]. The final

Table 1 Patient characteristics by individualised treatment effect on mortality tertiles in test set

	Lower third (predicted benefit from enhanced EM) *N* = 114	Middle third (predicted similar outcome w/ and w/o enhanced EM) *N* = 114	Upper third (predicted harm from enhanced EM) *N* = 114
Demographics, *n* (%)			
Age, years, mean (SD)	60 (17)	61 (14)	61 (14)
Body mass index, kg/m^2, mean (SD)	28.9 (7.6)	29.8 (7.2)	31.1 (7.0)
Female sex	46 (40.4%)	40 (35.1%)	36 (31.6%)
Employment status			
Retired or full-time home duties	56 (49.1%)	52 (45.6%)	49 (43.0%)
Unemployed or other[a]	13 (11.4%)	17 (14.9%)	23 (20.2%)
Employed, full or part-time[b]	45 (39.5%)	45 (39.5%)	42 (36.8%)
Type of home living			
Home alone or other	22 (19.3%)	24 (21.1%)	29 (25.4%)
Home living with other people	14 (12.3%)	17 (14.9%)	26 (22.8%)
Home with partner	78 (68.4%)	73 (64.0%)	59 (51.8%)
Frailty and function, median [IQR]			
Clinical frailty scale	3 [2, 4]	3 [2, 4]	3 [2, 4]
Functional Comorbidity Index	1 [1, 2]	2 [1, 3]	2 [1, 3]
Highest score on ICU mobility scale in week before ICU admission	10 [10, 10]	10 [10, 10]	10 [10, 10]
WHODAS 2.0	6.2 [0.0, 15.6]	10.4 [2.1, 22.9]	10.4 [2.1, 31.2]
Functional Comorbidity Index conditions, *n* (%)			
Anxiety, panic disorders, or depression	8 (7.0%)	16 (14.0%)	23 (20.2%)
Asthma	10 (8.8%)	15 (13.2%)	17 (14.9%)
Congestive heart failure or heart disease	25 (21.9%)	22 (19.3%)	19 (16.7%)
Diabetes type I and/or type II	6 (5.3%)	26 (22.8%)	45 (39.5%)
Upper gastrointestinal disease	12 (10.5%)	15 (13.2%)	11 (9.6%)
ICU admission, *n* (%)			
Hours from ICU admission to randomisation, median [IQR]	69.6 [43.8, 116.0]	50.2 [29.4, 85.6]	47.9 [25.6, 79.5]
Hours from hospital admission to randomisation, median [IQR]	108.3 [74.5, 171.8]	76.6 [50.3, 160.9]	65.3 [41.4, 120.2]
Diagnosis			
Non-operative			
Respiratory	30 (26.3%)	27 (23.7%)	18 (15.8%)
Sepsis	4 (3.5%)	16 (14.0%)	24 (21.1%)
Other	12 (10.5%)	16 (14.0%)	33 (28.9%)
Operative			
Cardiovascular	43 (37.7%)	29 (25.4%)	17 (14.9%)
Gastrointestinal	8 (7.0%)	5 (4.4%)	13 (11.4%)
Other	17 (14.9%)	21 (18.4%)	9 (7.9%)
Source			
Planned			
After elective surgery	56 (49.1%)	32 (28.1%)	10 (8.8%)
Unplanned			
From emergency surgery or from another hospital or ICU	19 (16.7%)	28 (24.6%)	62 (54.4%)
From emergency department	21 (18.4%)	31 (27.2%)	34 (29.8%)
From ward	18 (15.8%)	23 (20.2%)	8 (7.0%)
Pre-randomisation treatment, n (%)			
Corticosteroids	48 (42.1%)	51 (44.7%)	56 (49.1%)

Table 1 (continued)

	Lower third (predicted benefit from enhanced EM) *N* = 114	Middle third (predicted similar outcome w/ and w/o enhanced EM) *N* = 114	Upper third (predicted harm from enhanced EM) *N* = 114
Renal replacement therapy	24 (21.1%)	23 (20.2%)	33 (28.9%)
Vasopressor infusion	54 (47.4%)	76 (66.7%)	102 (89.5%)
Vitals and labs, median [IQR]			
APACHE II	15 [12, 19]	17 [13, 21]	18 [14, 24]
Creatinine	106 [73, 130]	107 [74, 158]	108 [78, 173]
GCS	15 [15, 15]	15 [15, 15]	15 [14, 15]
PEEP, mean (SD)	8.6 (2.9)	8.7 (3.1)	8.8 (2.8)
PF ratio, mean (SD)	247.8 (95.7)	234.8 (79.9)	242.7 (81.9)
PaO_2, mean (SD)	85.5 (20.5)	91.3 (45.9)	89.3 (27.1)
RASS	−2 [−4, −1]	−4 [−5, −2]	−4 [−5, −3]
Treatment, *n* (%)			
Early mobilisation	53 (46.5%)	58 (50.9%)	60 (52.6%)
Outcome, *n* (%)			
Death at day 180			
Total	18 (15.8%)	27 (23.7%)	32 (28.1%)
Early mobilisation	6 (11.3%)	13 (22.4%)	21 (35.0%)
Usual care	12 (19.7%)	14 (25.0%)	11 (20.4%)
Average treatment effect (ATE), % (95% CI)	−8.35% (−21.48, 4.77)	−2.59% (−18.20, 13.03)	14.63% (−1.53, 30.79)

EM early mobilisation, *SD* standard deviation, *IQR* interquartile range, *ICU* intensive care unit, *WHODAS* World Health Organisation Disability Assessment Schedule, *APACHE* acute physiology and chronic health evaluation, *GCS* Glasgow Coma Scale, *PEEP* positive end-expiratory pressure, *PF* PaO_2/FiO_2, *RASS* Richmond Agitation-Sedation Scale

[a] Unemployed or other = unemployed (due to health reasons), unemployed (due to other reasons), other

[b] Employed, full or part-time = employed, full or part-time, self-employed, full or part-time, student, full or part-time, non-paid work

model was then constructed using the algorithm with the highest discrimination (X-learner with a Bayesian additive regression tree or X-BART) built using all the training data. X-BART encompasses a meta-learner modelling strategy. The X-learner builds outcome models and subsequently pseudo-treatment effect models on treatment and control patients separately and then weights predictions obtained equally across models from both groups to form the final prediction. BART algorithms are a collection of decision trees and resemble gradient boosting machine algorithms in that each tree is built to improve the performance of the previous tree but with random changes implemented to the trees to reduce overfitting. Default hyperparameter settings from the original 'causaltoolbox' package were used [22]. Consistent with prior work, five distinct seed initialisations of the model were run, and the predictions were averaged across the runs to increase stability.

Model validation

Individualised treatment effect (ITE) predictions were obtained for each patient in the test set using the aggregation of the predictions from the X-BART models described above. The fundamental problem of causal inference is that the outcome for an individual under both the treated and the untreated states can never be simultaneously observed. Consequently, model performance assessment requires grouping similar patients and comparing treated versus untreated amongst the groups. Therefore, test dataset patients were ranked from most likely to benefit from EM (ITE closest to −1) to most likely to experience harm from EM (ITE closest to 1) and then divided into three equal-sized groups, similar to prior work [16]. Plotting the subgroup ATE for each third allowed for model assessment with a monotonic increase across thirds, indicating a model with high discrimination and calibration. The significance of these predictions was tested with a likelihood ratio test for the inclusion of the interaction term in a logistic regression model for the outcome with treatment and prediction as independent variables.

Discrimination was assessed further with the Qini coefficient [23] and C-for-benefit statistic [24]. The Qini coefficient is derived from the area under the Qini curve, which displays the cumulative uplift in 180-day survival observed when treating proportions of the population prioritising based on predicted ITE from the fitted model compared to random allocation. C-for-benefit quantifies

the probability of concordance between the predicted and observed benefit. A Qini coefficient above 0 and a C-for-benefit above 0.5 provide evidence of discrimination above that expected by chance [23, 24].

To investigate whether the differences in treatment response identified by the model were reflections of variation in treatment intensity, the amount of time (in minutes) spent at each ICU mobility scale (IMS) level was compared between the treatment groups across tertiles.

Finally, secondary length of stay outcomes were summarised across tertiles in patients with and without the 180-day mortality outcome.

Model explanation

Descriptive statistics were calculated for all predictor variables across quantiles to compare the characteristics of those predicted to benefit versus those predicted to experience harm at the group level. Median and interquartile range (IQR) or mean and standard deviation were calculated for continuous variables in accordance with the statistics reported in the original publication. For categorical variables, counts and percentages were reported.

To obtain a variable importance summary, we quantified the influence of each variable on the prediction for each patient in the training set by taking the absolute difference between the original prediction and the prediction obtained when replacing a predictor variable with the median value from the training data population with all else held constant [25].

Sensitivity analyses

Three sensitivity analyses were completed to assess the robustness of the results. First, the analysis was re-run with the train and test data swapped to ensure results were consistent with the primary analysis (eTable 3). Second, additional Qini coefficients with confidence intervals (CIs) were derived using augmented inverse propensity weights (AIPW) and bootstrapping to calculate 95% confidence intervals and test for the significance of the Qini coefficient in addition to the significance tests performed using the C-for-benefit and the likelihood ratio test. Third, tenfold cross-validation was performed across the full dataset, allowing for a larger sample size for training. Discrimination and the correlation of the out-of-sample predictions themselves were compared to the results reported in the primary analysis.

Results

Patient characteristics

The final analysis included 687 patients from 40 different sites, excluding patients from the UK because we did not have consent for secondary use of their data. There were 346 (50.4%) of the patients receiving EM and 141 (20.5%) who had died by day 180. As per the original trial, there was no significant difference between treatment groups for the outcome of 180-day mortality [ATE 95% CI 2.03% (−4.02, 8.08)]. After randomly splitting the data into train and test cohorts based on site, the training cohort contained 345 patients from 20 sites, and the test cohort contained 342 patients from 20 sites. A comparison of summary statistics for these two cohorts across predictor variables is shown in eTable 2. A higher proportion of the patients in the training set was admitted to the ICU from the ward and had a primary non-operative respiratory diagnosis at admission, whilst the patients in the testing set were admitted to the ICU after a planned surgery and had a primary operative cardiovascular diagnosis at admission with higher levels of PaO_2.

Model performance

In the training cohort, the X-BART model performed best and was used to develop the final model using all the training data. The predicted ITEs in the test cohort ranged from a 34.0 percentage point reduction to a 39.3 percentage point increase in mortality with enhanced EM. The observed ATEs across tertiles of predicted benefit from EM therapy are shown in Fig. 1. Amongst patients in the lower third (who were predicted to experience benefit from enhanced early mobility), the ATE was −8.35 (95% CI −21.48, 4.77). Patients in the middle third showed neither harm nor benefit from enhanced EM [ATE 95% CI −2.59% (−18.20, 13.03)]. Finally, patients in the upper tertile showed a point estimate indicating harm from enhanced EM [ATE 95% CI 14.63% (−1.53, 30.79)]. A significant likelihood ratio test ($p=0.006$) for the interaction between treatment and the ITE predictions indicated that the model's discrimination was significantly better than expected by chance. This was further corroborated by a Qini coefficient of 2.50 (eFig. 2), a C-for-benefit of 0.60 (0.53, 0.69), and as demonstrated in eFig. 3 (test set model calibration).

The mean differences in duration (minutes) spent at each IMS level did not differ significantly across tertiles, confirming the model was not identifying varying rates of treatment duration (eFig. 4).

Further, patients who experienced the outcome did not differ significantly in median days from randomisation to ICU discharge or randomisation to hospital discharge (eTable 4).

Model explanation

Values of the predictor variables across the tertiles can be seen in Table 1. Patients more likely to benefit from enhanced EM were characterised by lower

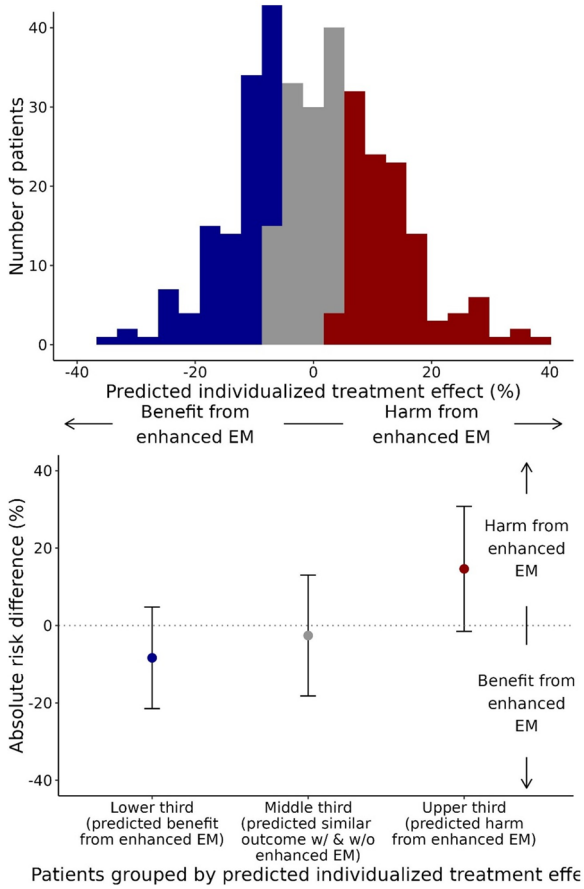

Fig. 1 Distribution of predicted individualised treatment effect distribution in the test set. The top plot shows a histogram of the distribution of predicted individualised treatment effects on mortality in the test set coloured according to the corresponding recommendation by third (benefit from EM, no benefit or harm from EM, harm from EM). The bottom plot shows the third on the *x*-axis and the absolute risk difference (aka average treatment effect) on the *y*-axis with 95% confidence intervals. A monotonic, increasing trend from left to right is indicative of treatment effect heterogeneity and high model discrimination

median World Health Organisation Disability Score (WHODAS, less disability) and functional comorbidity scores (fewer comorbidities), and a higher proportion of admissions from elective surgery compared to unplanned admissions. Patients more likely to experience harm from enhanced EM had lower median Richmond Agitation-Sedation Scale (RASS) scores, and a higher proportion were receiving vasopressors and diagnosed with diabetes at baseline. The top ten most important variables are demonstrated in Fig. 2A, and their trends can be seen in the individual conditional expectation (ICE) plot (eFig. 5). In the final model, the top three most important predictor variables were vasopressor infusion, RASS, and diabetes at baseline, and the relationship between these covariates and the ITE predictions is illustrated in Fig. 2B–D.

Sensitivity analyses

Three sensitivity analyses were performed. First, re-running the analysis run with the train and test data swapped resulted in a Qini coefficient of 2.14 (eTable 3) which demonstrates a similar level of discrimination to the primary analysis. Furthermore, diabetes, unplanned admission source "emergency surgery or from another hospital or ICU," Functional Comorbidity Index, and primary diagnosis of "other non-operative" and "sepsis non-operative" all showed up in the top ten variables for both models.

Second, for the original analysis of train versus test split, the AIPW Qini coefficient was 2.67 (95% CI 0.37–5.00) and for the analysis swapping of train and test data, the AIPW Qini coefficient was 2.40 (95% CI 0.21–4.59). Since these confidence intervals do not cross 0, they provide evidence that the model has discrimination that is statistically significantly better than chance.

Third, in the tenfold cross-validation using all the data, the median Qini coefficient in the left-out folds was 3.71, indicating slightly better performance than the 50/50 site split of the primary analysis. This increased performance is likely the result of the increased sample size for training and the random split being less rigorous than the site split. Furthermore, the correlation between predictions in the primary analysis test set and the tenfold cross-validation was high (Pearson's $r = 0.82$).

Discussion

This study sought to explore the ITEs of enhanced EM on critically ill patients in the ICU, with a specific focus on identifying which patient characteristics influence their response to the dosage of EM. Patients likely to benefit from enhanced EM were more likely to be admitted for planned surgery, and these patients were typically in better physical condition at baseline with less disability and fewer functional comorbidities. In contrast, patients who were predicted to experience harm from enhanced EM were more likely to have lower RASS scores at baseline, were more likely to require vasopressor support, and were more likely to have diabetes. These findings suggest that certain baseline factors, such as sedation level, need for vasopressor support, and comorbidities like diabetes, may impact a patient's ability to safely tolerate or benefit from enhanced EM.

We discovered significant HTE with EM in TEAM. Specifically, the predicted ITEs in the test cohort ranged from a 34.0 percentage point reduction to a 39.3

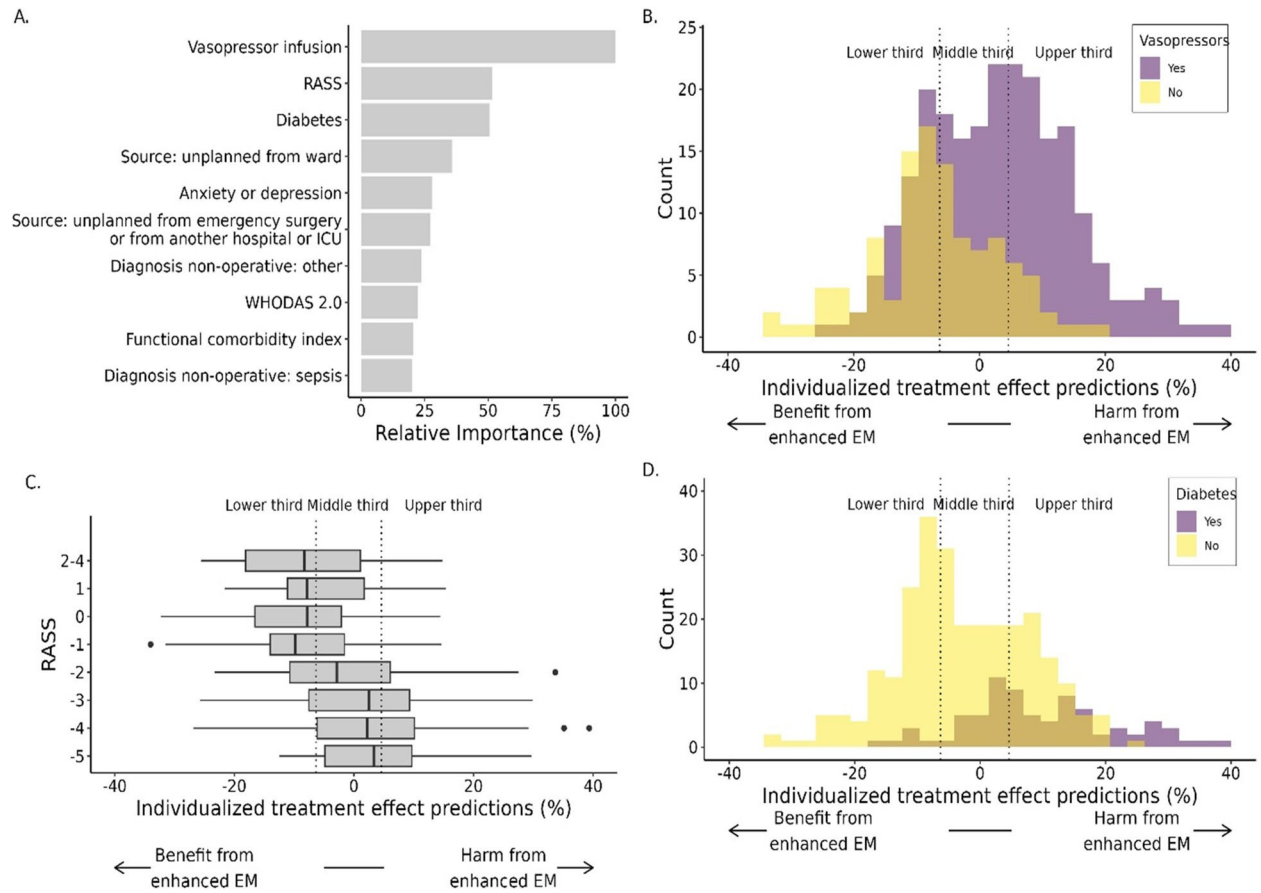

Fig. 2 Chart of relative variable importance and individual treatment effect predictions stratified by vasopressor status, RASS and diabetes status. In **A**, *x*-axis shows the relative importance and *y*-axis shows the top ten most important variables ordered from highest to lowest importance. The individual treatment effect predictions stratified by vasopressor status (**B**), RASS (**C**) and diabetes status (**D**). *EM* early mobilisation, *RASS* Richmond Agitation-Sedation Scale, *ICU* intensive care unit, *WHODAS* World Health Organisation Disability Assessment Schedule

percentage point increase in mortality with enhanced EM. These findings suggest that enhanced EM may be advantageous for some, but detrimental for others, although our confidence intervals were wide. The overall mortality rate at day 180, as reported in the original trial, did not show a significant difference between the treatment groups, but our individualised analysis reveals that there may be important baseline variables within the patient population that can be used to predict individualised estimates of benefit or harm for mortality from enhanced EM, which were not captured by the ATE. Given the low number of adverse events and their previously published lack of association with mortality, further stratified AE analysis would be underpowered and unlikely to yield additional insight. We have, therefore, focussed on mortality-based heterogeneity.

The three most influential variables identified by the model, vasopressor use, diabetes, and RASS score, are clinically plausible predictors of response to EM. Previous secondary analyses of the TEAM trial suggested increased mortality associated with diabetes [26, 27]. The current analysis indicates that whilst many patients with diabetes are harmed with enhanced EM, a small number of patients with diabetes may benefit from enhanced EM. This individual treatment response to enhanced mobilisation in patients with diabetes may be due to the difference in severity of complications. Patients with underlying microvascular disease and poor perfusion, requiring vasopressors, or with autonomic dysfunction resulting in hypotension or arrhythmias, may have increased mortality. Alternatively, patients with fewer complications may respond well to enhanced mobilisation with improved glycaemic control, reduced ICU-acquired weakness, or improved cardiovascular function, resulting in improved survival. Patients receiving vasopressors and those with low RASS scores may be at increased risk of harm from EM due to their underlying physiological instability, haemodynamic compromise, and reduced

capacity for active participation. Mobilising patients in this state may exacerbate hypotension, increase myocardial oxygen demand, and potentially lead to adverse events such as arrhythmias. Importantly, the study's findings align with previous research emphasising the need for individualised treatment strategies in critical care settings [14, 27, 28]. In a previous study, healthy patients responded well to enhanced EM, whilst patients with pre-existing chronic disease required targeted EM mapped to their premorbid status [29]. Individually tailored sessions of mobilisation during the TEAM trial were based primarily on moment-to-moment clinical status, safety parameters, muscle strength and the patient's immediate tolerance of activity. In contrast, when we refer to the HTE analysis in this study for *individualised rehabilitation*, we are highlighting a broader concept. Here, individualised rehabilitation refers to the potential for longer term personalisation of *dosage, intensity and trajectory* of rehabilitation based on patient-level characteristics (e.g. comorbidity burden, sedation levels, organ support). The HTEs revealed in this analysis highlight the limitations of a one-size-fits-all approach in EM in ICU. This reinforces the importance of a tailored approach to mobilisation, where patient characteristics help guide the decision-making process.

The timing of EM in ICU has been a key focus of research, with several small studies in medical ICU populations suggesting benefits from initiating mobilisation as early as possible after ICU admission [7, 11]. In the TEAM trial, patients were enrolled a median of 60 h after ICU admission, and 80% received a physiotherapy treatment on the day of randomisation. Our findings indicate that patients from mixed ICUs (including medical, surgical, and trauma units) who benefit from enhanced EM mostly had a longer interval between ICU admission and randomisation compared to those who experienced harm. This suggests that the optimal timing of mobilisation may vary based on individual patient characteristics and the trajectory of critical illness. The timing of treatment exposure, specifically the interval between ICU admission and randomisation, may play a critical role in modifying the effects of EM. Some patients randomised earlier in the ICU stay may be more physiologically unstable, with higher sedation levels, vasopressor requirements, or ongoing organ support, making them less likely to tolerate or benefit from enhanced mobilisation. This temporal variability in treatment exposure could influence both the feasibility and effectiveness of EM and may partially explain the observed heterogeneity in ITEs. Future analyses should consider incorporating time-to-randomisation as a covariate or stratification factor to better understand its interaction with treatment response.

A key strength of this study was the use of machine learning to assess treatment effect heterogeneity. Traditional analyses often rely on fixed subgroup analyses that fail to consider the complexity of individual patient characteristics. For example, the secondary analysis of the TEAM study found that patients with diabetes had increased mortality with higher dose mobilisation compared to lower dose mobilisation, but it failed to show harm in other patients without diabetes [27]. It also failed to show HTEs in patients with diabetes, and this analysis indicates that some patients with diabetes may benefit from enhanced EM. Indeed, this is the first study to identify a potential benefit in patients with enhanced EM. Prior post hoc analyses have only identified harm or no effect [27]. Machine learning techniques, such as X-BART, offer a more comprehensive approach by evaluating multiple factors simultaneously and creating individualised predictions of benefit. The sensitivity analysis, which involved swapping the train and test datasets, reinforced the robustness of these findings, demonstrating that the results were consistent even with different data splits. The Qini coefficient and C-for-benefit statistic provide insight into the model's ability to discriminate and rank patients by their predicted benefit from treatment. The **Qini coefficient** quantifies the cumulative gain in outcomes (e.g. survival) when patients are prioritised for treatment based on predicted ITEs, compared to random allocation. A Qini coefficient greater than zero indicates that the model provides value in identifying patients who benefit most compared to random allocation. In this analysis, a Qini coefficient of **2.50** suggests meaningful discrimination, with a similar value (**2.14**) observed in the sensitivity analysis, supporting model robustness. The **C-for-benefit statistic** measures concordance between predicted and observed treatment benefit, analogous to the C-statistic in risk prediction models. A value of **0.60** (95% CI 0.53–0.69) indicates moderate discrimination, exceeding the threshold of 0.5 expected by chance. Whilst benchmarks for these metrics vary by context, values above 0.5 for C-for-benefit and positive Qini coefficients are generally considered indicative of potentially useful predictive performance in treatment effect heterogeneity analyses.

Whilst the study provides valuable insights into the individualised effects of EM, there are limitations to consider. The ITE estimates derived from the X-BART model are framed as causal effects, but valid causal interpretation requires several key assumptions. First, exchangeability must hold, meaning that—conditional on observed covariates—patients assigned to enhanced EM and usual care EM are comparable. This is

supported by the randomised design of the TEAM trial, which helps mitigate confounding. Second, positivity assumes that all patients have a non-zero probability of receiving either treatment, which is generally satisfied in RCTs but may be challenged by site-specific practices or exclusion criteria. Third, consistency requires that the treatment received corresponds precisely to the treatment defined in the analysis, which may be affected by variations in EM implementation across sites. Fourth, correct model specification is critical in machine learning-based causal inference; misspecification or overfitting can bias ITE estimates. Although the use of cross-validation, a held-out test set, and sensitivity analyses enhances robustness, the complexity of the model and the secondary nature of the analysis warrant caution. These assumptions, particularly in the context of flexible machine learning algorithms, should be considered when interpreting the ITE estimates as causal effects.

It is important to acknowledge that the original TEAM trial reported no overall mortality benefit from enhanced mobilisation. According to the PATH Statement [20], predictive HTE analyses conducted in the context of null trials are particularly susceptible to spurious findings unless they are strongly justified a priori. Whilst our analysis was pre-specified and followed rigorous methodological standards, including external validation and sensitivity analyses, the exploratory nature of this secondary analysis means that the findings should be interpreted with caution. Because absolute risk reductions (ARR) can vary with baseline risk, our use of ARR to summarise subgroup effects may reflect underlying differences in risk rather than true heterogeneity of treatment efficacy; relative effect measures such as risk ratios may offer more consistent comparisons across subgroups and should be considered in future analyses. An additional limitation to note is that enhanced EM was treated as a binary intervention to preserve the integrity of randomisation; this approach does not account for dose, timing or duration of actual exposure, which may have influenced observed treatment effects and limits inferences about individual-level responsiveness. Also, the post hoc exclusion of UK data, although necessary due to lack of consent, may have introduced selection bias, reduced statistical power and limited the generalisability of findings across different international healthcare settings. Although 180-day mortality was selected to align with the original TEAM trial, we recognise that outcomes during this longer time frame may be influenced by post-ICU co-interventions not captured in our analysis. Finally, whilst the model demonstrates strong predictive performance, the clinical applicability requires careful consideration, and further prospective studies are needed to validate these findings in the clinical setting. Future work could incorporate bootstrap-based confidence intervals for ITEs to further quantify model uncertainty.

Conclusion

This hypothesis-generating analysis supports further exploration of the individualised effects of the dosage of EM in critically ill ICU patients. By leveraging machine learning to explore treatment effect heterogeneity, key patient characteristics were identified that may influence the response to the dosage of EM. These findings advocate for a more personalised approach to mobilisation in the ICU, where patient-specific factors may guide the decision of whether and when to deliver enhanced EM interventions. Future studies should explore the clinical applicability of these individualised predictions and assess whether they should be integrated into routine ICU care to optimise patient outcomes.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1007/s00134-025-08217-0.

Author details
[1] Australian and New Zealand Intensive Care Research Centre, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia. [2] Department of Critical Care, University of Melbourne, Melbourne, Australia. [3] Intensive Care Unit and Physiotherapy Department, The Alfred Hospital, Melbourne, Australia. [4] Critical Care Division, The George Institute for Global Health, Sydney, Australia. [5] Division of Pulmonary and Critical Care, Department of Medicine, University of Wisconsin-Madison, Madison, USA. [6] Data Analytics Research and Evaluation Centre, Austin Health, Melbourne, Australia. [7] Department of Intensive Care, Royal Melbourne Hospital, Melbourne, Australia. [8] University College Dublin-Clinical Research Centre at St. Vincent's University Hospital, Dublin, Ireland. [9] Intensive Care Service, Royal Prince Alfred Hospital, Sydney Local Health District, Sydney, Australia. [10] School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia. [11] Clinical Trials Unit, Intensive Care National Audit and Research Centre (ICNARC), London, UK. [12] Department of Physiotherapy, Royal Perth Hospital, Royal Perth Bentley Group, East Metropolitan Health Service, Perth, Australia. [13] Curtin School of Allied Health, Curtin University, Bentley, Australia. [14] Medical Research Institute of New Zealand, Wellington, New Zealand. [15] Johns Hopkins University, Baltimore, USA. [16] Department of Intensive Care, Alfred Hospital, Melbourne, Australia. [17] School of Medicine, University of Melbourne, Parkville, Melbourne, Australia. [18] Department of Anesthesiology and Intensive Care Medicine (CCM/CVK), Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin, Germany. [19] Clinical Division of General Anaesthesia and Intensive Care Medicine, Department of Anaesthesia, Intensive Care Medicine and Pain Medicine, Medical University of Vienna, Vienna, Austria. [20] Intensive Care Unit, St. John of God Hospital Subiaco, Perth, Australia. [21] Division of Clinical and Translational Research, Washington University School of Medicine, St Louis, USA. [22] Division of Critical Care, Department of Anesthesia, Washington University School of Medicine, St Louis, USA. [23] Division of Allergy, Pulmonary, and Critical Care Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, USA. [24] Vanderbilt Institute for Clinical and Translational Research, Nashville, USA. [25] Intensive Care Unit, Wellington Hospital, Wellington, New Zealand.

Author contributions
All authors contributed to the writing of the manuscript and agreed on the final version.

Funding
The TEAM trial was supported by a NHMRC Project Grant (GNT1120319) from Australia led by CLH. CLH is supported by a NHMRC Investigator Grant (GNT2033103) from Australia. RB was supported by a NHMRC Investigator Grant (GNT2033013) from Australia. AMH is supported by a NHMRC Investigator Grant (GNT2008447) from Australia. MWS is supported, in part, by a grant from the NIH/NCATS (5UL1TR002243). The funding agencies had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data availability
Data sharing will only be considered and approved by the TEAM Study Steering Committee after written request.

Declarations

Conflicts of interest
The authors declared no relevant conflicts of interest with respect to the research, authorship and/or publication of this article. CLH and SJS are Section Editors for Intensive Care Medicine. They have not taken part in the review or selection process of this article.

Ethics approval
The TEAM trial was approved by the institutional review boards of each participating country, and a waiver of informed consent was approved for the data to be transferred to the University of Wisconsin–Madison for analysis.

Publisher's Note

Received: 9 July 2025 Accepted: 12 November 2025
Published: 24 November 2025

References

1. Fan E, Dowdy DW, Colantuoni E, Mendez-Tellez PA, Sevransky JE, Shanholtz C, Himmelfarb CR, Desai SV, Ciesla N, Herridge MS, Pronovost PJ, Needham DM (2014) Physical complications in acute lung injury survivors: a two-year longitudinal prospective study. Crit Care Med 42:849–859
2. Puthucheary ZA, Rawal J, McPhail M, Connolly B, Ratnayake G, Chan P, Hopkinson NS, Phadke R, Dew T, Sidhu PS, Velloso C, Seymour J, Agley CC, Selby A, Limb M, Edwards LM, Smith K, Rowlerson A, Rennie MJ, Moxham J, Harridge SD, Hart N, Montgomery HE (2013) Acute skeletal muscle wasting in critical illness. JAMA 310:1591–1600
3. Lad H, Saumur TM, Herridge MS, Dos Santos CC, Mathur S, Batt J, Gilbert PM (2020) Intensive care unit-acquired weakness: not just another muscle atrophying condition. Int J Mol Sci 21:7840
4. Hermans G, Van Mechelen H, Clerckx B, Vanhullebusch T, Mesotten D, Wilmer A, Casaer MP, Meersseman P, Debaveye Y, Van Cromphaut S, Wouters PJ, Gosselink R, Van den Berghe G (2014) Acute outcomes and 1-year mortality of intensive care unit-acquired weakness. A cohort study and propensity-matched analysis. Am J Respir Crit Care Med 190:410–420
5. Van Aerde N, Meersseman P, Debaveye Y, Wilmer A, Gunst J, Casaer MP, Bruyninckx F, Wouters PJ, Gosselink R, Van den Berghe G, Hermans G (2020) Five-year impact of ICU-acquired neuromuscular complications: a prospective, observational study. Intensive Care Med 46:1184–1193
6. Schaller SJ, Anstey M, Blobner M, Edrich T, Grabitz SD, Gradwohl-Matis I, Heim M, Houle T, Kurth T, Latronico N, Lee J, Meyer MJ, Peponis T, Talmor D, Velmahos GC, Waak K, Walz JM, Zafonte R, Eikermann M, International Early S-gMRI (2016) Early, goal-directed mobilisation in the surgical intensive care unit: a randomised controlled trial. Lancet 388:1377–1388
7. Schweickert WD, Pohlman MC, Pohlman AS, Nigos C, Pawlik AJ, Esbrook CL, Spears L, Miller M, Franczyk M, Deprizio D, Schmidt GA, Bowman A, Barr R, McCallister KE, Hall JB, Kress JP (2009) Early physical and occupational therapy in mechanically ventilated, critically ill patients: a randomised controlled trial. Lancet 373:1874–1882
8. Devlin JW, Skrobik Y, Gelinas C, Needham DM, Slooter AJC, Pandharipande PP, Watson PL, Weinhouse GL, Nunnally ME, Rochwerg B, Balas MC, van den Boogaard M, Bosma KJ, Brummel NE, Chanques G, Denehy L, Drouot X, Fraser GL, Harris JE, Joffe AM, Kho ME, Kress JP, Lanphere JA, McKinley S, Neufeld KJ, Pisani MA, Payen JF, Pun BT, Puntillo KA, Riker RR, Robinson BRH, Shehabi Y, Szumita PM, Winkelman C, Centofanti JE, Price C, Nikayin S, Misak CJ, Flood PD, Kiedrowski K, Alhazzani W (2018) Clinical practice guidelines for the prevention and management of pain, agitation/sedation, delirium, immobility, and sleep disruption in adult patients in the ICU. Crit Care Med 46:e825–e873
9. Lang JK, Paykel MS, Haines KJ, Hodgson CL (2020) Clinical practice guidelines for early mobilization in the ICU: a systematic review. Crit Care Med 48:e1121–e1128
10. Paton M, Chan S, Tipping CJ, Stratton A, Serpa Neto A, Lane R, Young PJ, Romero L, Broadley T, Hodgson CL (2023) The effect of mobilization at 6 months after critical illness—meta-analysis. NEJM Evid 2:EVIDoa2200234
11. Patel BK, Wolfe KS, Patel SB, Dugan KC, Esbrook CL, Pawlik AJ, Stulberg M, Kemple C, Teele M, Zeleny E, Hedeker D, Pohlman AS, Arora VM, Hall JB, Kress JP (2023) Effect of early mobilisation on long-term cognitive impairment in critical illness in the USA: a randomised controlled trial. Lancet Respir Med 11:563–572
12. Iwashyna TA-O, Burke JF, Sussman JB, Prescott HC, Hayward RA, Angus DC (2015) Implications of heterogeneity of treatment effect for reporting and analysis of randomized trials in critical care. Am J Respir Crit Care Med 192:1045–1051
13. Broadley T, Serpa Neto A, Bailey M, Bellomo R, Brickell K, Buhr H, Gabbe BJ, Gould DW, Harrold M, Hurford S, Iwashyna TJ, Nichol AD, Presneill JJ, Schaller SJ, Sivasuthan J, Tipping CJ, Webb S, Young PJ, Higgins AM, Hodgson CL, TEAM Study Investigators (2025) Adverse events during and after early mobilisation: A post hoc analysis of the TEAM trial. Aust Crit Care 38:101156
14. Fuest KE, Ulm B, Daum N, Lindholz M, Lorenz M, Blobner K, Langer N, Hodgson C, Herridge M, Blobner M, Schaller SJ (2023) Clustering of critically ill patients using an individualized learning approach enables dose optimization of mobilization in the ICU. Crit Care 27:1
15. Curth A, Peck RW, McKinney E, Weatherall J, van der Schaar M (2024) Using machine learning to individualize treatment effect estimation: challenges and opportunities. Clin Pharmacol Ther 115:710–719
16. Buell KG, Spicer AB, Casey JD, Seitz KP, Qian ET, Graham Linck EJ, Self WH, Rice TW, Sinha P, Young PJ, Semler MW, Churpek MM (2024) Individualized treatment effects of oxygen targets in mechanically ventilated critically ill adults. JAMA 331:1195–1204
17. Seitz KP, Spicer AB, Casey JD, Buell KG, Qian ET, Graham Linck EJ, Driver BE, Self WH, Ginde AA, Trent SA, Gandotra S, Smith LM, Page DB, Vonderhaar DJ, West JR, Joffe AM, Doerschug KC, Hughes CG, Whitson MR, Prekker ME, Rice TW, Sinha P, Semler MW, Churpek MM (2023) Individualized treatment effects of bougie versus stylet for tracheal intubation in critical illness. Am J Respir Crit Care Med 207:1602–1611

18. Jaki T, Chang C, Kuhlemeier A, Van Horn ML (2024) Predicting Individual Treatment Effects: Challenges and Opportunities for Machine Learning and Artificial Intelligence. KI Künstliche Intell 39:27–32
19. TEAM Study Investigators, the ANZICS CTG, Hodgson CL, Bailey M, Bellomo R, Brickell K, Broadley T, Buhr H, Gabbe BJ, Gould DW, Harrold M, Higgins AM, Hurford S, Iwashyna TJ, Serpa Neto A, Nichol AD, Presneill JJ, Schaller SJ, Sivasuthan J, Tipping CJ, Webb S, Young PJ (2022) Early active mobilization during mechanical ventilation in the ICU. N Engl J Med 387:1747–1758
20. Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, Ioannidis JPA, Patrick-Lake B, Morton S, Pencina M, Raman G, Ross JS, Selker HP, Varadhan R, Vickers A, Wong JB, Steyerberg EW (2020) The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann Intern Med 172:35–45
21. Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28:1–26
22. Kunzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci USA 116:4156–4165
23. Yadlowsky S, Fleming S, Shah N, Brunskill E, Wager S (2025) Evaluating treatment prioritization rules via rank-weighted average treatment effects. J Am Stat Assoc 120:38–51
24. van Klaveren D, Steyerberg EW, Serruys PW, Kent DM (2018) The proposed "concordance-statistic for benefit" provided a useful metric when modeling heterogeneous treatment effects. J Clin Epidemiol 94:59–68
25. Molnar C (2020) Interpretable machine learning. Leanpub
26. Lindholz M, Schellenberg CM, Grunow JJ, Kagerbauer S, Milnik A, Zickler D, Angermair S, Reisshauer A, Witzenrath M, Menk M, Boie S, Balzer F, Schaller SJ (2022) Mobilisation of critically ill patients receiving norepinephrine: a retrospective cohort study. Crit Care 26:362
27. Serpa Neto A, Bailey M, Seller D, Agli A, Bellomo R, Brickell K, Broadley T, Buhr H, Gabbe BJ, Gould DW, Harrold M, Higgins AM, Hurford S, Iwashyna TJ, Nichol AD, Presneill JJ, Schaller SJ, Sivasuthan J, Tipping CJ, Poole A, Parke R, Bradley S, Webb S, Zoungas S, Young PJ, Hodgson CL, TEAM Study Investigators (2024) Impact of high-dose early mobilization on outcomes for patients with diabetes: a secondary analysis of the TEAM trial. Am J Respir Crit Care Med 210:779–787
28. Schaller SJ, Scheffenbichler FT, Bein T, Blobner M, Grunow JJ, Hamsen U, Hermes C, Kaltwasser A, Lewald H, Nydahl P, Reisshauer A, Renzewitz L, Siemon K, Staudinger T, Ullrich R, Weber-Carstens S, Wrigge H, Zergiebel D, Coldewey SM (2024) Guideline on positioning and early mobilisation in the critically ill by an expert panel. Intensive Care Med 50:1211–1227
29. Puthucheary ZA, Denehy L (2015) Exercise interventions in critical illness survivors: understanding inclusion and stratification criteria. Am J Respir Crit Care Med 191:1464–1467